Analyzing Work Sample Task Performance Using Three Data Sets

Donald L. Harville
Armstrong Laboratory Human Resources Directorate
Brooks AFB, Texas 78235-5352

Abstract

The Armed Services Vocational Aptitude Battery (ASVAB) has historically been validated against technical school grades. Job performance measurement (JPM) data were collected in the 1980s to link ASVAB cut scores to more realistic job performance measures for eight Air Force Specialties (AFSs). Due to the expense of these work samples, a major challenge has been to find reliable and valid surrogates. For 261 Aerospace Ground Equipment Mechanics (423x5) first-termers, data on performing 16 hands-on tasks were collected. These tasks, selected by subject matter experts (SMEs) at special JPM workshops, ranged from having seven to 26 steps and from low to high difficulty. Three data sets were generated for each task -- members getting all, some, or none of the task's steps correct. For all tasks, the largest numbers of personnel were associated with getting some steps correct. Generally, less-difficult tasks were associated with larger numbers of personnel getting all steps correct. Getting all steps correct and some steps correct tended to be close in average task performance time, as well as having smaller average times, than did getting no steps correct. A correlation of .79 ($p < .001$) was found between the average time for all steps correct and independent SME estimates of the average time required by first-termers for task completion. Members with all steps correct tended to have more recent task experience, more overall task experience, more average task performance per month, higher task experience ratings, higher Armed Forces Qualification Test (AFQT) scores, and higher grades in technical school. Future research needs are discussed.

Introduction

The United States Air Force (USAF) has historically validated the *g*-loaded (Ree & Earles, 1992) Armed Services Vocational Aptitude Battery (ASVAB) against grades in technical school. ASVAB scores are used as part of the process of USAF enlisted selection and classification. While grades in technical school are important, assuming that the content of technical school training and testing accurately reflects the content of the relevant Air Force Specialty (AFS), they do not cover the entire job performance criterion space. Because of this lack of coverage and a problem with the norming of the ASVAB, in 1980 the Office of the Secretary of Defense started a Joint-Service program to research the Job Performance Measurement (JPM) of enlisted personnel. The resulting performance measures were to be used to directly link enlistment standards to job performance. JPM data were collected for eight AFSs over a period of three years in the mid 1980s, and have been analyzed using numerous methodologies. The current study analyzed these data in a new way for Aerospace Ground Equipment (AGE) Mechanics (AFS 423x5). AGE was of one of the last four AFSs that data were collected on. The current analyses used three data sets for each task -- members getting all, some, or none of the task's steps correct. Armed Forces Qualification Test (AFQT) scores, final technical school grades, four task experience measures, and task difficulty were examined as predictors of task completion. One purpose of the study was to examine how well each of these measures predicted membership in each of the three data sets for each task. Successful task completion times for getting all of the task's steps correct were also compared against independent expert estimates of the average task completion times required by all first-termers for successful task performance.

Method

The examinees were 261 USAF AGE enlisted incumbents (240 males, 21 females; 229 Whites, 20 Blacks, 12 other) in their first four years of enlistment (Laue, Bentley, Bierstedt, & Molina, 1992). The mean experience for the examinees was 28.4 months. A work sample of 16 AGE tasks was tested using hands-on performance testing to assess the job performance of the incumbents on tasks representative of their Air Force Specialty (AFS). The 16 tasks were chosen partially on the basis of job analyses of their frequency of performance and difficulty. Prior to task selection as part of routine occupational inventory surveys by the Air Force Occupational Measurement Squadron, task difficulties were provided by subject matter experts (SMEs) using a 9-point scale. Higher task difficulty numbers referred to more difficult tasks. Another set of SMEs was used to select tasks that represented each AFS and could be tested, given equipment and testing time constraints. These SMEs also defined the tasks in terms of the steps involved in successful task completion and the equipment used. For purposes of task selection, more difficult tasks were given a priority compared to less difficult tasks.

The work sample tests required the incumbents to perform the tasks while being observed by trained test administrators. These extensively trained test administrators were provided information from SMEs concerning

task objective, task time limits, and tools and equipment required to perform the task. The examinees were allowed access to routinely used technical documents and were instructed to perform each task according to normal procedures. If they asked why they were being timed, examinees were told that their task times were being collected only for administrative purposes (i.e., in order to finish all the scheduled testing in the allotted time). For scoring purposes, each task was divided into the steps necessary for successful task completion. While an examinee performed a task, the test administrator marked on a checklist whether or not each step was correctly performed. If a subject reached the maximum time limit for a task, all uncompleted steps were recorded as incorrect. The 16 tasks ranged from having seven steps for a task, "Research technical orders for AGE chassis, enclosure, and drive maintenance information" to 26 steps, "Perform a gas turbine compressor inspection." Four task-level experience measures were used: 1) weeks since last performed; 2) number of times performed; 3) average times performed per month (calculated by dividing number of times performed by job experience); and 4) task experience ratings using a 5-point scale. Of the five variables mentioned above, only job experience was not self-reported.

The Armed Forces Qualification Test (AFQT) was used as the aptitude measure and is comprised of four subtests of the ASVAB. The ASVAB has 10 ability subtests (M = 50, SD = 10) with a range from 20 to 80 points. The sum of two of these scores (Arithmetic Reasoning and Mathematical Knowledge) plus twice the Verbal score (a sum of Word Knowledge and Paragraph Comprehension) comprises the AFQT.

Average task completion times and Ns were determined for three data sets for each of the 16 tasks -- members getting all, some, or none of the task's steps correct. Independent SME estimates of average successful task completion times for all first-termers were compared against the actual average times. Differences among the three data sets in the four task experience measures, AFQT scores, and technical school final grades are reported.

Results

As reported in Table 1, for all 16 tasks the largest numbers of personnel were associated with getting some steps correct. Less-difficult tasks were associated with larger numbers of personnel getting all steps correct ($r = -.567$, $p = .022$). Across the 16 tasks, the average percentage of personnel with all steps correct was 7.5%, some steps correct 85.2%, and no steps correct 7.3%. For one task, "Remove or install hydraulic lines or fittings," all personnel got only some of the steps correct. This task had only eight steps, a mid-level task difficulty of 4.86, and mid-level time limit of 15 minutes. For an additional task, "Perform gas turbine compressor inspections," no personnel got all steps correct. This was the second most difficult task and had 26 steps. It also had the longest time limit (45 minutes).

A correlation of .790 ($p = .0008$) was found between the average times for all steps correct (N=261) and independent SME estimates of the average time required by all first-termers for successful task completion.

Table 1. Summary task information using three data sets for each task

(Task difficulty) Task description (Task time limit in min.)	All steps correct	Some steps correct	No steps correct
	Avg. time in min. (N)	Avg. time in min. (N)	Avg. time in min. (N)
(5.08) Perform AGE electrical checks (20)	12.8 (16)	12.1 (205)	19.5 (40)
(4.49) Perform load bank service inspections (12)	8 (2)	6.8 (233)	9 (3)
(6.47) Adjust turbine engine fuel system components (20)	17 (5)	16.6 (133)	19.9 (123)
(5.31) Measure resistance of AGE electrical components (20)	11.8 (4)	9.9 (252)	20 (3)
(5.13) Perform generator service inspections (15)	8.8 (33)	8.5 (228)	N/A (0)
(4.59) Research tech orders, charts, or diagrams (12)	6.9 (19)	6.3 (235)	10.5 (6)
(4.06) Splice electrical system wiring (25)	14.7 (31)	15.0 (229)	25 (1)
(4.09) Remove or install fuel lines or fittings (12)	5.6 (30)	5.4 (231)	N/A (0)
(4.76) Clean motor or generator armature (15)	8.7 (32)	9.1 (191)	14.9 (36)
(6.03) Isolate engine, motor, or generator malfunctions (35)	22.5 (2)	21.1 (258)	35 (1)
(6.25) Perform gas turbine compressor inspections (45)	N/A (0)	39.2 (237)	45 (24)
(5.74) Perform hydraulic test stand service inspections (15)	9.5 (22)	10.4 (239)	N/A (0)
(4.86) Remove or install hydraulic lines or fittings (15)	N/A (0)	6.4 (261)	N/A (0)
(3.59) Remove and replace engine fan belts (25)	8.1 (98)	7.5 (161)	2.5 (2)
(5.83) Isolate pneumatic system malfunctions (25)	8.8 (6)	17.1 (187)	24.8 (65)
(3.71) Inspect vehicles for safety of operations (12)	9.1 (9)	9.6 (252)	N/A (0)

Figure 1. Subject matter expert estimates of average task times required by all first-termers for successful task completion versus average times for all steps correct (N=261) presents a scatterplot for these data points. The

most variation between the estimated and actual averages was for the tasks estimated by the experts as requiring 15 minutes.

Numerous small cell sizes makes analyses of variance problematic for the three data sets for each task. All steps correct and some steps correct tended to be close in average task performance time, as well as having smaller average times than no steps correct. For 10 tasks, the longest task times were associated with no steps correct. As reported in Table 2, using tasks in the same order as in Table 1, members with all steps correct usually had the most recent task experience, most overall task experience, most average task performance per month, highest task experience ratings, highest Armed Forces Qualification Test (AFQT) scores, and highest grades in technical school. On the average, those examinees who were correct on all steps for the first task in the table had performed it 5.6 months ago, those correct on no steps had performed it 15.6 months ago, while those correct on some steps had performed it 16.1 months ago. Therefore those examinees with most recent average task experience, the smallest average time since last performing this task, had all steps correct. Table 2 reflects this by having E1, the symbol used for the most recent mean task experience, in the column for all steps correct for the first task. For seven tasks, members with all steps correct had the most recent task experience. For nine tasks, members with all steps correct had the most overall task experience. For eight tasks, members with all steps correct had the most task experience per month. For 11 tasks, members with all steps correct had the highest task experience ratings. For 11 tasks, members with all steps correct had the highest AFQT scores. For 13 tasks, members with all steps correct had the highest final technical school grades. Members with no steps correct tended to be at the other extreme on these six measures, while members with some steps correct tended to be in the middle.

Table 2. Partial rank orderings for means of four experience measures, AFQT, and final school grade within each task

TD	All steps correct	Some steps correct	No steps correct
5.08	E1, E4, AFQT, FSG	E2, E3	
4.49	E2, E3, E4, FSG	E1, AFQT	
6.47	AFQT, FSG	E2, E3, E4	E1
5.31	E2, E3, E4, AFQT, FSG		E1
5.13	E1, E2, E3, E4, AFQT, FSG		N/A
4.59	E2, E3, E4, AFQT, FSG		E1
4.06	E1, E2, E3, E4, AFQT		FSG
4.09	AFQT, FSG	E1, E2, E3, E4	N/A
4.76	E1, E2, E3, E4, AFQT, FSG		
6.03	E1, E4, FSG	E2, E3	AFQT
6.25	N/A	E1, E2, E3, E4, AFQT	FSG
5.74	E1, E2, E3, E4, AFQT, FSG		N/A
4.86	N/A	N/A	N/A
3.59	E2, FSG	E1, E3	AFQT
5.83	E1, E4, AFQT, FSG	E2, E3	
3.71	E2, E3, E4, AFQT, FSG	E1	N/A

Note. TD = task difficulty for the task; E1 = most recent mean task experience; E2 = most overall mean task experience; E3 = most mean task experience per month; E4 = highest mean task experience ratings; AFQT = highest mean AFQT score; FSG = highest mean final school grade. Experience ratings were not collected for the task with a TD of 3.59.

Discussion

The most interesting result was the surprisingly large .79 correlation between SME estimates of the average times required for successful task performance by first-termers and the results from the study for all steps correct. This indicates that SME time estimates may be useful for job/AFS restructuring. For such purposes, the SMEs would need to agree on the steps and equipment involved in performing a task, as they did during workshops conducted for this study prior to data collection.

Data collected for the JPM project cost approximately one million dollars for each AFS, or $5000 for each airman in the sample. It is doubtful that anything similar to this massive amount of carefully conceptualized work sample data will be collected and available for use in the near future. Therefore, it is important that the lessons learned from this project be applied to all relevant, smaller work sample data-collection efforts. These lessons include careful standardization of tasks, having the instructions to the examinees always stating why they are

being timed, timing the steps in addition to the complete tasks, and perhaps even videotaping task performance if Privacy Act concerns can be addressed. Having times at the step level in addition to the task level could have important implications for the training communities, in addition to the manpower and classification communities.

References

Laue, F.J., Bentley, B.A., Bierstedt, S.A., & Molina, R. (1992). Data collection and administration procedures for the Job Performance Measurement system (AL-TR-1992-0118). Brooks AFB, TX: Human Resources Directorate.

Ree, M.J., & Earles, J.A. (1992). Subtest and composite validity of ASVAB Forms 11, 12, and 13 for technical training courses (AL-TR-1991-0107). Brooks AFB, TX: Human Resources Directorate.

Back to Table of Contents

INTERNET DOCUMENT INFORMATION FORM

A . Report Title: Analyzing Work Sample Task Performance Using Three Data Sets

B. DATE Report Downloaded From the Internet: 06/04/99

C. Report's Point of Contact: (Name, Organization, Address, Office Symbol, & Ph #): **Navy Advancement Center**
ATTN: Dr. Grover Diel (850) 452-1615
Pensacola, FL

D. Currently Applicable Classification Level: Unclassified

E. Distribution Statement A: Approved for Public Release

F. The foregoing information was compiled and provided by:
DTIC-OCA, Initials: __VM__ **Preparation Date 06/04/99**

The foregoing information should exactly correspond to the Title, Report Number, and the Date on the accompanying report document. If there are mismatches, or other questions, contact the above OCA Representative for resolution.